**CS 134: Networks** — **Problem set 3**

Prof. Yaron Singer — Due: **Wednesday 2/22/2017**

**1. Non-Preferential Attachment (20 points).** Consider the following model of non-preferential attachment networks:

- At time $t = 1$, there is one node (node 1) in the network.
- At time $t > 1$, node $t$ is added and it establishes one directed edge from node $t$ to $i_1, \cdots, i_{t-1}$ where each $i_j$ is picked independently and uniformly at random between 1 and $t - 1$.

**a. (8 points)** At time step $t$, what is the expected in-degree[1] of node $i_k, k \in \{1, \ldots, t-1\}$?

**b. (8 points)** At time step $t$, what is the fraction of nodes that have expected in-degree at least $d$?

**c. (4 points)** Do you expect graphs generated using this model to have a power-law degree distribution? Briefly explain why or why not.

**Note:** In this problem, your answers should be closed-form formulas (e.g., they should not contain a summation from 1 to $t$). You may treat the following approximation as if it were exact: $\sum_{i=1}^{n} \frac{1}{i} \approx \log(n)$.

**Solution: (Thanks, Andrew Zhou!)**

[1] The in-degree of a node $u$ is the number of edges directed towards $u$.

**a. (8 points)** At time step $t$, what is the expected in-degree[1] of node $i_k, k \in \{1, \ldots, t-1\}$?

Let us consider node $i_1$. At time step 1 the expected degree is 1 because the degree is guaranteed to be 1. At time step 2, there is an 100% chance that node $i_2$ will connect to $i_1$, so the expected degree increases by 1. At time step 3, there is a 50% chance that node $i_3$ will connect to $i_1$ (the other 50% is for $i_2$) so the expected degree increases by $\frac{1}{2}$. We may repeat this logic to conclude that at time step $t$, there is a $\frac{1}{t-1}$ chance that the new node $i_k$ will connect to $i_1$. Thus the expected in-degree of node $i_1$ at time step $t$ is $\frac{1}{1}+\frac{1}{2}+...+\frac{1}{t-1} = \sum_{i=1}^{t-1}\frac{1}{i} = \log(t-1)$. (This is true for $t > 1$; for $t = 1$ the expected degree is 0.)

We now note that for node 2, we have the same series of possible connections, save that the there is no chance to connect to node $i_2$ at time step 1. Similarly, for node $i_k$, we have the same series of connections except for time steps 1 to $k$ there is no chance of connecting to $i_k$. So we have that the expected in-degree of $i_k$ at time step $t > k$ is $\frac{1}{k}+...+\frac{1}{t-1} = \sum_{i=1}^{t-1}\frac{1}{i} - \sum_{i=1}^{k-1}\frac{1}{i} = \log(t-1) - \log(k-1)$. The expected degree at $t \leq k$ is 0.

**b. (8 points)** At time step $t$, what is the fraction of nodes that have expected in-degree at least $d$?

From above we have that the expected in-degree of node $i_k$ at time step $t$ is $\log(t-1)-\log(k-1)$. Let us solve for $j$ such that the expected in-degree of all nodes $i_k$, $k \leq j$ have expected in-degree at least $d$.

$$\begin{aligned}
\log(t-1) - \log(k-1) &\geq d \\
\log(\frac{t-1}{k-1}) &\geq d \\
\frac{t-1}{k-1} &\geq e^d \\
k-1 &\leq (t-1)e^{-d} \\
k &\leq (t-1)e^{-d}+1
\end{aligned}$$

So we have that for $k \leq (t-1)e^{-d}+1$, $i_k$ has expected in-degree at least $d$ at time step $t$. At time step $t$ there are $t$ nodes in the network, so the fraction is $\frac{(t-1)e^{-d}+1}{t}$. This translates as $P(d \geq k) = \frac{(t-1)e^{-d}+1}{t}$ supposing we sample uniformly from the $t$ nodes in the network at time step $t$.

**c. (4 points)** Do you expect graphs generated using this model to have a power-law degree distribution? Briefly explain why or why not.

We have $P(d \geq k) = ck^{-\alpha}$ for a power-law degree distribution. At time step $t$ we are considering nodes $i_1$ to $i_t$. We have that $P(d \geq k) = \frac{(t-1)e^{-d}+1}{t} = \frac{t-1}{t}e^{-d} + \frac{1}{t}$. Regarding $t$ as a constant and disregarding the term $\frac{1}{t}$, we have $P(d \geq k) = ce^{-d}$, which is exponential rather than a power-law degree distribution. We would not expect graphs generated with this model to follow the power-law degree distribution.

**2. Power Laws in Classrooms [Easley and Kleinberg, 18.8 Q3] (10 points).** Suppose that some researchers studying educational institutions decide to collect data to address the following two questions.

- As a function of $k$, what fraction of Harvard classes have $k$ students enrolled?
- As a function of $k$, what fraction of third-grade elementary school classrooms in New York

State have $k$ pupils?

Which one of these would you expect to more closely follow a power-law distribution as a function of $k$? Give a brief explanation for your answer, using some of the ideas about power-law distributions.

**Solution: (Thanks, Michelle Vaccaro!)**

**Solution:** I would expect the fraction of Harvard classes that have $k$ students enrolled to more closely follow a power law distribution as a function of $k$ because the quantity being measured can be seen as a measure of popularity, namely the popularity of the class, since students at Harvard choose their classes whereas third graders cannot. Because the fraction of Harvard classes that have $k$ students enrolled is a measure of the popularity of the class, we expect to see extreme imbalances, with very large values likely to arise. Indeed, we see this occur in Harvard classes like CS50, Ec10a/10b, LS1a/1b, Hebrew Bible, etc. These large values account for the heavy tail we see in the power law distribution. In contrast, since third grade elementary school classrooms in New York State have a state mandated maximum number of students, these extreme imbalances will not arise and there will not be a heavy tail.

**3. It's a Power-Law Story (15 points).** Suppose for \$100 you can purchase the exclusive rights to sell the latest single by Taylor Swift. However, due to branding restrictions you can only sell the single for \$1 to each buyer. You are now trying to decide whether or not purchasing the rights for \$100 would be financially wise. Of course, you don't know how many people will actually purchase the song, which we can model as a random variable $X$. After doing some research, you discover that the likelihood of a single having exactly $k$ purchases follows a power-law distribution with $\alpha = 2$ for all values of $k \in \{1, \cdots, 485165195\}$, with the likelihood of any more purchases being zero (you read a recent survey in the Economist tipping you off that there are exactly 485,165,195 people in the world who are willing to pay for music; the rest follow their conscience and try to get it for free). Assume that your expected profit is your expected revenue from selling the single minus the cost for the sales rights (in this case, \$100). If you're trying to make a positive profit, should you purchase the rights?

**Solution by Lars Lorch:** We are given a random variable $X$, denoting the number of sales of the song, with a support set $k \in \{1, ..., 485165195\}$ following a power-law distribution with $\alpha = 2$. Any probability of sales higher than 485165195 songs has $p = 0$. $X$ is discrete. Thus, $c$ in $P(X = x) = c * x^{-2}$ needs to be defined by the following condition.

$$\sum_{x=1}^{485165195} c * x^{-2} = 1$$

$$c = \frac{1}{\sum_{x=1}^{485165195} x^{-2}} \approx \frac{1}{1.64493}$$

$$c \approx 0.60793$$

We can now properly define our distribution:

$$P(X = x) = 0.60793 * x^{-2}$$

Since we get \$1 for every song sold, our expected return is the same as $E[X]$.

$$E[X] = \sum_{x=1}^{485165195} x * 0.60793 * x^{-2}$$

$$\approx 12.5095$$

No, if I'm trying to make positive profits and the rights cost \$100, I should not buy the rights.

**4. Power-Law Cliques (25 points).** Fix a graph $G = (V, E)$. Construct a graph $\widetilde{G}$ as follows. For every node $i \in V$ with degree $k$, let $\widetilde{G}$ contain a clique $C_i$ of $k + 1$ nodes. These cliques are all distinct, and $\widetilde{G}$ is the union of all of them.

**a. (3 points)** Suppose $V = \{a, b, c, d\}$ and $E = \{ab, bc, ac, ad\}$. Draw $\widetilde{G}$.

**b. (7 points)** For an arbitrary $G$, let $p$ be the degree distribution of $G$, so that $p(k)$ is the fraction of nodes with degree $k$ in $G$. Denote by $\Delta$ the maximum degree in $G$. Let $\widetilde{p}$ denote the degree distribution of $\widetilde{G}$. Give a formula for $\widetilde{p}(k)$ in terms of $p$ (i.e., in terms of the values of $p$).

**c. (8 points)** We say a distribution $q$ on the natural numbers has the *power-law property until* $K$ if there is a constant $\beta > 1$ so that

$$\frac{q(k')}{q(k)} \geq \left(\frac{k'}{k}\right)^{-\beta}$$

whenever $k$ and $k'$ are integers such that $1 \leq k \leq k' \leq K$. You can check that this is consistent with the definition of a power-law distribution from lecture.

Suppose that $p$ has the power-law property until $\Delta$. Show that $\widetilde{p}$ has the power-law property until $\Delta$ as well (possibly with a different constant $\beta$).

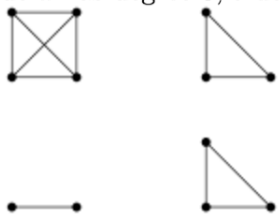
**d. (7 points)** The "strong form" of the friendship paradox says that for a typical person $i$, the average degree of $i$'s neighbors will strictly exceed $i$'s own degree. Does this necessarily hold in networks whose degree distributions have the power law property?

**Solution: (Thanks, Stephen Yen!)**

## 4.

### a.

Node $a$ has degree 3, $b$ degree 2, $c$ degree 2, and $d$ degree 1. So, $\tilde{G}$ can be represented:



### b.

For each node of degree $k$ in $G$, there are $k+1$ nodes with degree $k$ in $\tilde{G}$. We can analyze the possible clique generated by one node to find $\tilde{p}(k)$:

$$\tilde{p}(k) = \frac{(k+1)p(k)}{\sum_{j=1}^{\Delta}(j+1)p(j)}$$

This expression is the average number of nodes with degree $k$ generated by a node over the average number of nodes generated.

### c.

If $p$ has the power law property up to $\Delta$, then for $k \in [1, \Delta]$

$$p(k) = ck^{-\alpha}$$

Then, for $1 \leq k \leq k' \leq K$,

$$\frac{p(k')}{p(k)} = \frac{ck'^{-\alpha}}{ck^{-\alpha}} = \frac{k'^{-\alpha}}{k^{-\alpha}} = (\frac{k'}{k})^{-\alpha}$$

We can let $\beta = c\alpha$ for some $c \geq 1$. Then, since $k' \geq k$,

$$\frac{p(k')}{p(k)} = (\frac{k'}{k})^{-\alpha} \geq (\frac{k'}{k})^{-c\alpha} = (\frac{k'}{k})^{-\beta}$$

as desired.

**d.**

This statement is false. Consider the graph $\tilde{G}$ described in 4a. $k \in \{2, 3, 4\}$, and

$$p(1) = \frac{1}{6}$$

$$p(2) = \frac{1}{2}$$

$$p(3) = \frac{1}{3}$$

Letting $B = 2$, we can see that the power law property on $K$ for $K = 3$ holds:

$$\frac{p(2)}{p(1)} = 3 \geq 2^{-2}$$

$$\frac{p(3)}{p(2)} = \frac{3}{2} \geq \frac{3}{2}^{-2}$$

However, because all of the nodes belong to cliques, each node has exactly the same number of neighbors as its neighbors. Thus, by counterexample, the power law property does not imply that the strong form of the friendship paradox holds.

**5. Coding: Fitting Power-Laws (30 points).** In this coding problem, we will use the files network1.txt and network2.txt in the pset folder. In each file, the two numbers per line represent an edge from the first node to the second. For all plotting, we advise you to use matplotlib's pyplot function.

**a. (5 points)** For each data set, plot the degree distribution of the social network with the degree $d$ on the horizontal axis and the fraction of nodes with degree $d$ on the vertical axis, i.e., $f(d) = \frac{|\{v \in V : \text{v has degree } d\}|}{|V|}$.

**b. (5 points)** For each data set, plot the degree distribution of the graph in log-log scale. That is, for $d$ and $f(d)$ as above, plot $\log d$ on the horizontal axis and $\log f(d)$ on the vertical axis. Does the plot seem linear?

**c. (7 points)** Using your $\log - \log$ plot, fit the following simplified power-law model using ordinary least squares regression (see section notes for details) with an appropriately chosen bin size for : $\log p(x) = \alpha \log x + b$ where $b$ is some constant.

**d. (10 points)** For each data set, fit the power-law model to the data sets using the method of maximum likelihood. Specifically, write a method which, given an array of degrees (where the $i$th entry of this array is the degree of node $i$), fits the parameter $\alpha$ of the power law model to the given input.

**e. (3 points)** Plot the power law distributions for your estimates for parameters from parts (c) and (d). Briefly discuss the qualitative differences between the two methods, why these

difference might happen, and the appropriateness of a power law model for these datasets.